BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

Detail 75(1- 150)

Publication No. : 1020030072709 (20030919)
Application No. : 1020020011871 (20020306)
Title of Invention : METHOD OF SELECTING PROBE SET FOR GENOTYPING
Document Code : A
IPC : C12Q 1/68
Priority :
Applicant : SAMSUNG ELECTRONICS CO., LTD.
Inventor : SIM, HYEONG SEOP

Abstract :

PURPOSE: Provided is a method of selecting a genotyping probe set which identifies the difference between the normal target nucleic acid and the mutation target nucleic acid, and even in the case of insertion/deletion, makes the genotyping with a few probes.

CONSTITUTION: A method of selecting a genotyping probe set comprises the steps of: hybridizing the normal target nucleic acid and the mutation target nucleic acid, and collecting the hybridization intensity data of the probes thereof; testing the average differences of the probes based on the data, and excepting the probe with no significant difference; computing the false positive error rate where the normal target nucleic acid is falsely classified to be the mutation target one or the false negative error rate where the mutation target nucleic acid is falsely classified to be the normal target one by making a cross validation to the probes; and selecting the probe passing the standard of the false positive or negative error rate.

Legal Status :

1. Appliaction for a patent (20020306)

(19)대한민국특허청(KR)
(12) 공개특허공보(A)

(51) 。 Int. Cl.7
C12Q 1/68

(11) 공개번호 특2003-0072709
(43) 공개일자 2003년09월19일

(21) 출원번호 10-2002-0011871
(22) 출원일자 2002년03월06일

(71) 출원인 삼성전자주식회사
경기도 수원시 팔달구 매탄3동 416번지

(72) 발명자 심형섭
경기도수원시권선구권선동유원아파트603동1205호

(74) 대리인 이영필
이해영

심사청구 : 있음

(54) 유전자형 확인용 프로브 세트 선택 방법

요약

본 발명은 정상 표적 핵산에 상보적인 복수 개의 프로브(probe) 및 돌연변이 표적 핵산에 상보적인 복수개의 프로브가 고정된 바이오칩 상에서 정상 표적 핵산 및 돌연변이 표적 핵산을 혼성화 (hybridization) 시켜 정상 표적 핵산-프로브 및 돌연변이 표적 핵산-프로브의 혼성화 강도(hybridization intensity) 데이터를 수집하는 단계; 상기의 데이터를 이용하여 평균차 검정을 실시하여 유의한 차이가 나지 않는 프로브를 제거하는 단계; 유의한 차이를 가지는 프로브의 정상 표적 핵산-프로브 혼성화 강도 분포 및 돌연변이 표적 핵산-프로브 혼성화 강도 분포로부터 정상 표적 핵산을 돌연변이 표적으로 분류할 확률인 가짜 양성 에러율(false positive error rate) 또는 돌연변이 표적을 정상 표적 핵산으로 잘못 분류할 확률인 가짜 음성 에러율(false negative error rate)을 계산하는 단계; 및 상기 가짜 양성 에러율 또는 가짜 음성 에러율의 기준을 통과한 프로브를 선택하는 단계를 포함하는 유전자형 확인용 프로브 세트 선택 방법에 관한 것이다.

본 발명에 따르면, 정상 표적 핵산과 돌연변이 표적 핵산의 차이를 실험 오차 이내에서 감지할 수 있고, 점돌연변이 (point mutation) 뿐만 아니라 삽입(insertion) / 결실(deletion)의 경우에도 소수의 프로브를 가지고 유전자형을 확인할 수 있다.

대표도

도 3

명세서

도면의 간단한 설명

도 1은 본 발명의 동작원리를 나타내는 흐름도이다.

도 2는 혼성화 실험에 의한 데이타 수집 단계의 흐름도이다.

도 3은 정상 표적 핵산과 프로브와의 혼성화 강도 분포와, 돌연변이 표적 핵산과 프로브와의 혼성화 강도 분포에

대한 평균차 검정을 실시하여 프로브를 심사하고, 유의하고 올바른 프로브만을 선발하는 단계의 흐름도이다.

도 4는 돌연변이 위치 3306의 두 돌연변이형 프로브(MP1, MP2)와 정상형/돌연변이형 표적과의 혼성화 강도 분포를 나타낸 산점도이다.

도 5는 돌연변이 위치 4037의 두 돌연변이형 프로브(MP1, MP2)와 정상형/돌연변이형 표적과의 혼성화 강도 분포를 나타낸 산점도이다.

도 6은 돌연변이 위치 5683의 두 돌연변이형 프로브(MP1, MP2)와 정상형/돌연변이형 표적과의 혼성화 강도 분포를 나타낸 산점도이다.

도 7은 돌연변이 위치 6195의 두 돌연변이형 프로브(MP1, MP2)와 정상형/돌연변이형 표적과의 혼성화 강도 분포를 나타낸 산점도이다.

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

본 발명은 유전자형을 모르는 샘플이 정상 유전자를 가졌는지 아니면 돌연변이형 유전자를 가졌는지를 확인하는 바이오칩에 고정될 프로브 세트를 선택하는 방법에 관한 것이다. 구체적으로 in silico 방법에 의해 설계된 프로브 풀(pool)이 고정된 바이오칩과 유전자형을 이미 알고 있는 샘플(sample)과의 혼성화 실험결과를 통한 데이터 수집 단계, 평균차 검정을 통해 프로브를 심사 및 선발하는 단계, 다양한 프로브 조합에 의한 판별성능 평가 단계, 및 최종 프로브 세트를 선택하는 단계를 포함하는 유전자형 확인에 필요한 프로브 세트를 선택하는 방법에 관한 것이다.

본 발명에 의해 선택된 프로브는 바이오칩 실험에서 발생하는 다양한 오차를 감안하여 선택되기 때문에 실제로 최적의 판별 성능을 보이고, 본 발명에 의해 선택된 프로브를 탑재한 바이오칩은 유전자형을 최소의 오차로 정확하게 판별하기 때문에 유전자의 돌연변이와 질병과의 연관(association)을 연구하는데 활용할 수 있다.

유전자형을 확인하기 위한 방법에는 제한효소로 처리된 단편을 크기별로 분리하는 방법(restriction fragment sizing), 대립형질 특이성 올리고뉴클레오타이드 혼성화(allele specific oligonucleotide hybridization), 변성 구배 겔 전기영동(denaturing gradient gel electrophoresis), 단일 가닥 구조 분석(single stranded conformation analysis) 등 여러가지가 있지만 최근에 개발된 방법인 DNA 칩의 경우는 동시에 여러 위치의 유전자형을 한꺼번에 판별할 수 있는 장점이 있 어 많은 사람들의 관심을 끌고 있다.

DNA 칩의 선두주자인 Affymetrix는 올리고뉴클레오타이드(oligonucleotide)를 프로브로 하는 DNA 칩을 사용하여 돌연변이(미국 특허 제6,027,880호, Arrays of nucleic acid probes and methods of using the same for detecting cystic fibrosis) 또는 다형성(미국 특허 제6,300,063호, Polymorphysm detection)등의 염기서열의 돌연변이를 파악하는 방법을 이미 제시하였다.

Affymetrix의 유전자형 검출 방법은 기본적으로는 돌연변이가 있다고 알려진 위치에 A,C,G,T를 가지고 돌연변이가 있지 않은 다른 부위는 모두 같은 염기배열을 가지는 길이 9~25mer의 올리고뉴클레오타이드(oligonucleotide)를 프로브로 사용하는 타일드 어레이(tiled array) 방식이다. 이 방법은 서열결정(sequencing)의 기능을 함께 달성하기 위해서 돌연변이가 있다고 알려진 염기 서열의 위치뿐만 아니라 돌연변이 위치의 근처에서도 타일드 어레이 방식으로 가능한 모든 조합의 염기서열을 프로브로 사용하기 때문에 타일드 어레이를 적용하는 위치의 수가 증가함에 따라 필요한 프로브의 숫자도 4배씩 증가한다.

이에, 본 발명자는 돌연변이 위치가 정확하게 알려진 염기서열에 대해서 최소의 프로브를 사용하면서도 정확한 유전자형을 판별할 수 있는 방법을 찾기 위하여 연구 노력한 결과, 혼성화를 통한 데이터 수집 단계, 프로브의 심사 및 선발 단계, 성능평가 단계, 및 최종 프로브 세트를 선택하는 단계를 통하여 유전자형 확인에 필요한 프로브 세트를 선택하는 경우, 정상 표적 핵산과 돌연변이 표적 핵산의 차이를 실험 오차 이내에서 감지할 수 있고, 점돌연변이 (point mutation) 뿐만 아 니라 삽입(insertion) / 결실(deletion)의 경우에도 소수의 프로브를 가지고 유전자형을 확인할 수 있음을 확인하고, 본 발명을 완성하게 되었다.

발명이 이루고자 하는 기술적 과제

따라서, 본 발명의 주된 목적은 정상 표적 핵산과 돌연변이 표적 핵산의 차이를 실험 오차 이내에서 감지할 수 있고, 점돌연변이 (point mutation) 뿐만 아니라 삽입(insertion) / 결실(deletion)의 경우에도 소수의 프로브를 가지고 유전자형을 확인할 수 있는 유전자형 확인용 프로브 세트 선택 방법을 제공하는 데 있다.

발명의 구성 및 작용

본 발명의 목적을 달성하기 위하여, 정상 표적 핵산에 상보적인 복수 개의 프로브 및 돌연변이 표적 핵산에 상보적인 복수 개의 프로브가 고정된 바이오칩과 정상 표적 핵산을 가지고 있는 샘플 및 돌연변이 표적 핵산을 가지고 있는 샘플을 반복적으로 혼성화 (hybridization) 시켜 정상 표적 핵산을 가지고 있는 샘플-프로브 (본 프로브는 정상형 및 돌연변이형이 혼합된 프로브이다) 및 돌연변이 표적 핵산을 가지고 있는 샘플-프로브(본 프로브도 정상형 및 돌연변이형이 혼합된 프로브이다)의 혼성화 강도 데이터를 수집하는 단계; 상기의 데이터를 이용하여 평균차 검정(t-test)을 실시하여 유의한 차이가 나지 않는(샘플이 다른데도 불구하고 혼성화 강도의 평균이 통계적으로 다르지 않은 것으로 p 값이 유의수준보다 큰 것을 의미) 프로브를 제외하는 프로브 심사 및 선발단계; 유의한 차이를 가지는 프로브(p 값이 유의수준보다 작은 프로브)에 대하여 정상 표적 핵산-프로브 혼성화 강도 분포 및 돌연변이 표적 핵산-프로브 혼성화 강도 분포로부터, 교차 유효성 검증법(cross-validation)에 의해 정상 표적 핵산을 돌연변이 표적으로 분류할 확률인 가짜 양성 에러율(false positive error rate)과 돌연변이 표적을 정상 표적 핵산으로 잘못 분류할 확률인 가짜 음성 에러율(false negative error rate)을 계산하는 프로브 성능평가 단계; 및 상기 가짜 양성 에러율이 기준보다 낮고 동시에 가짜 음성 에러율의 기준보다 낮은 프로브를 선택하는 단계를 포함하는 유전자형 확인용 프로브 세트 선택 방법을 제공한다.

본 발명의 프로브 심사 및 선발단계에 있어서, 두 표본(정상 표적 핵산을 가지는 샘플-프로브와의 혼성화 강도 분포, 돌연변이 표적 핵산을 가지는 샘플-프로브와의 혼성화 강도 분포)에 대해서 모두 정규분포를 따르면 두 표본 t-검정(two sample t-test)을 실시하여 p 값을 구하고, 어느 하나라도 정규분포를 따르지 않는 경우는 비모수적 방법(nonparametric method)에 의해 p 값을 구하는 단계; 상기 p 값으로 두 표본의 평균이 유의한 차이가 있는가를 확인하는 단계(유의한 차이가 난다는 것은 유의 수준(α)보다 p 값이 작은 것을 의미한다); 상기 유의한 차이가 있는가를 확인하는 단계에서 유의한 차이가 나는 프로브에 대해 정상형 프로브의 경우는 정상 표적 핵산과의 혼성화 강도 평균이 돌연변이 표적 핵산과의 혼성화 강도 평균보다 크고 돌연변이형 프로브는 반대로 돌연변이 표적 핵산과의 혼성화 강도 평균이 정상 표적 핵산과의 혼성화 강도 평균보다 큰 프로브를 선발하는 단계를 포함할 수 있다.

본 발명의 두 표본에서 평균에 대한 t-검정을 실시하는 단계에서, 두 표본에서 평균에 대한 t-검정을 실시하기 전에 각 샘플에 대한 혼성화 강도의 분산이 같은지를 검정(등분산성 검정)하여 같은 경우는 t-검정에서 나온 결과 중에 등분산에 해당하는 p 값을 선택하고, 다른 경우는 t-검정에서 나온 결과 중에 등분산이 아닌 경우에 해당하는 p 값을 선택하는 단계를 더 포함할 수 있다.

본 발명의 p 값으로 두 표본의 평균이 유의한 차이가 있는가를 확인하는 단계에서, 유의수준은 0.01, 바람직하게는 0.001로 정한다. p값이 유의 수준보다 작으면 유의한 차이가 있다고 판정한다.

이하, 본 발명을 단계별로 보다 구체적으로 설명한다.

본 발명은 혼성화 강도 데이터 수집(hybridization intensity collection) 단계, 프로브 심사 및 선발(probe screening) 단계, 프로브 성능평가(probe quality estimation) 단계, 최종 프로브 세트를 선택하는 단계를 거쳐 유전자형 확인 (genotyping)에 필요한 프로브 세트를 선택한다.

<u>혼성화 강도 데이터 수집 단계</u>

본 과정을 도2에 도시하였으므로 참조하시오. in silico 방법에 의해 실험 계획된 프로브들은 상보적인 샘플의 서열(sequence), 올리고뉴클레오타이드의 길이와 유전자에서 유전자형 확인(genotyping)을 하려고 하는 부분의 위치에 따라 각각 다르다. 정상 표적 핵산에 상보적인 복수의 프로브(이를 '정상형 프로브: wild type probes'라 함)와 돌연변이 표적 핵산에 상보적인 복수의 프로브(이를 '돌연변이형 프로브:mutant type probes'라 함)를 칩 상에 고정시키고, 정상 표적 핵산을 가진 샘플 및 돌연변이 표적 핵산을 가진 샘플과 혼성화(hybridization)시킨다. 한 장의 DNA 칩 상에 <u>배열된</u> 모든 프로브의 위치를 알고 있기 때문에 혼성화를 마치고 건조된 칩에서 스캐너를 이용하여 각 프로브들과 유전자형을 알고 있는 샘플과의 혼성화 강도를 얻게 된다. 동일한 프로브가 심겨진 여러 장의 DNA칩에서 이 과정을 반복해서 각 프로브별로 데이터를 정리하면 정상 표적 핵산을 가진 샘플-프로브의 혼성화 강도 분포와 돌연변이 표적 핵산을 가진 샘플-프로브의 혼성화 강도 분포를 얻게 된다.

프로브 심사 및 선발(screening) 단계

본 과정을 도3에 도시하였으므로 참조하시오. 각 프로브 별로 정리된 혼성화 강도 데이터에서 정상 표적 핵산을 가지는 샘플-프로브의 혼성화 강도(I_w)의 평균과 표준 편차를 각각 μ_w, σ_w로 표시하고, 돌연변이표적 핵산을 가지는 샘플-프로브의 혼성화 강도(I_m)의 평균과 표준 편차를 각각 μ_m, σ_m로 표시한다. 각 샘플에 대해 모든 칩에서 얻어진 혼성화 강도를 표준화한 절대값($|(I-\mu)/\sigma|$)이 3을 넘는 경우가 있는지를 확인하여, 3을 넘는 경우를 이상치(outlier)라고 말한다. 이상치를 가지는 경우 실험오차가 아닌지 확인한 후에 실험오차로 판명되면 그 데이타는 제거한다.

정상 표적 핵산-프로브와의 혼성화 강도 평균(μ_w)과 돌연변이 표적 핵산-프로브와의 혼성화 강도 평균(μ_m)이 통계적으로 유의한 차이가 나는지를 확인하기 위하여 평균차 검정을 실시한다. 먼저 각 샘플-프로브와의 혼성화 강도의 분포가 정규분포를 따르는지 확인한다. 두 표본(정상 표적 핵산을 가지는 샘플-프로브와의 혼성화 강도 분포, 돌연변이 표적 핵산을 가지는 샘플-프로브와의 혼성화 강도 분 포)에 대해서 모두 정규분포를 따르면 두 표본 t-검정(two sample t-test)을 실시하여 p 값을 얻고, 어느 하나라도 정규분포를 따르지 않는 경우는 비모수적 방법(nonparametric method)에 의해 p 값을 얻어서 두 표본의 평균이 유의한 차이가 있는지를 확인한다. 유의한 차이가 난다는 것은 우리가 잡은 유의 수준보다 p 값의 값이 작은 것을 의미한다. 두 표본에서 평균에 대한 t-검정을 실시하기 전에 각 샘플에 대한 혼성화 강도의 분산이 같은지를 검정(등분산성 검정)하여 같은 경우는 t-검정에서 나온 결과 중에 등분산에 해당하는 p 값을 선택하고, 다른 경우는 t-검정에서 나온 결과 중에 등분산이 아닌 경우에 해당하는 p 값을 선택한다. 유의한 차이가 나는 프로브(이를 '유효한 프로브:valid probe'라 한다)에 대해서는 정상형 프로브의 경우는 정상 표적 핵산과의 혼성화 강도 평균이 돌연변이 표적 핵산과의 혼성화 강도 평균보다 커야만 하고 돌연변이형 프로브는 반대로 돌연변이 표적 핵산과의 혼성화 강도 평균이 정상 표적 핵산과의 혼성화 강도 평균보다 커야 한다는 조건(이것을 도3에서 'I(PM)>I(MM)'이라고 표현했다. I는 혼성화 강도(intensity), PM은 정상형 프로브-정상 표적 핵산, 돌연변이형 프로브-돌연변이 표적 핵산으로 결합한 경우(완전일치:perfect match), MM는 정상형 프로브-돌연변이형 표적 핵산, 돌연변이형 프로브-정성형 표적 핵산으로 결합한 경우(불일치:mismatch)를 가리킨다)을 만족하는지 확인 후 만족하는 프로브(이를 '올바른 프로브: right probe'라 하고, 만족하지 못한 것을 '잘못된 프로브: wrong probe'라 한다)만 선발한다. 위에서 설명한 평균차 검정 과정을 탑재된 모든 프로브에 대해서 반복한다.

프로브 성능평가(estimation) 단계

프로브 심사 및 선발(probe screening) 단계에서 정상 표적 핵산-프로브와의 혼성화 강도 평균(μ_w)과 돌연변이 표적 핵산-프로브와의 혼성화 강도 평균(μ_m)이 같다는 평균차 검정의 귀무가설($H_0: \mu_w = \mu_m$)을 기각(p 값 <σ)한 프로브들만 성능평가를 실시한다. 한 돌연변이 위치(mutation site)의 유전자형을 판별하기 위한 복수의 프로브 풀들 중에서 p 값이 가장 작은 프로브의 성능을 제일 먼저 평가하고 p 값이 낮은 순서로 성능을 평가한다. 한 프로브가 여러 개의 칩에 대해 반복한 실험을 통해 정상 표적 핵산과의 혼성화에 의해 얻어진 강도 값들과($I^1, I^2, \ldots, I^k$)과 돌연변이 표적 핵산과의 혼성화에 의해 얻어진 강도 값들($J^1, J^2, \ldots, J^m$) 중에서 I^1을 빼고 로지스틱 회귀분석(logistic regression)을 실시해서 얻어진 판별함수에 I^1의 값을 대입했을 때 정상으로 판별하면 오차가 없는 것이고 돌연변이로 판별하면 가짜 양성 에러를 범하는 것이다. k번째 혼성화 강도까지 위와 같은 작업을 하고 나서 가짜 음성 에러를 범하는 회수를 k로 나누면 가짜 양성 에러율(false positive error rate)을 얻는다. 그리고 나서 J^1을 빼고 로지스틱 회귀분석을 실시해서 얻어진 판별함수에 J^1의 값을 대입했을 때 돌연변이로 판별하면 오차가 없는 것이고 정상으로 판별하면 가짜 음성 에러를 범하는 것이다. 마찬가지로 m번째 혼성화 강도까지 위의 작업을 반복해서 가짜 음성 에러(false negative error)를 범하는 회수를 m으로 나누면 가짜 음성 에러율(false negative error rate)을 얻는다. 이와 같은 과정을 교차 유효성 검증법(cross-validation)이라고 한다. 위에서 설명한 방법으로 평균차 검정에서 유의한 프로브들에 대해 모두 교차 유효성 검증법을 실시한다.

프로브 세트 선택 단계

유전자형을 판별하고자 하는 각 위치에 대해서 프로브 성능평가 단계에서 얻어진 가짜 양성 에러율(false positive error rate)과 가짜 음성 에러율(false negative error rate)에 적당한 가중치(weight)를 곱해서 얻어지는 값이 작은 순서대로 정렬(sorting)해서 프로브를 선택한다. 예를 들어, 진단의 입장에서는 가짜 양성 에러율이 가짜 음성 에러율보다 많이 중요하다고 하면 가짜 양성 에러율에 대한 가중치를 3으로 하고 가짜 음성 에러율에 대한 가중치로는 1을 둘 수 있다. 최종적인 판별 함수는 이렇게 구해진 2개 이상의 프로브들을 가지고 구성한다. 이 판별함수에 유전자형을 모르는 샘플과 혼성화했을 때 나오는 혼성화 강도 값을 대입하여 유전자형을 결정한다.

이하, 실시예를 통하여 본 발명을 더욱 상세히 설명하기로 한다. 이들 실시예는 단지 본 발명을 예시하기 위한 것이므로, 본 발명의 범위가 이들 실시예에 의해 제한되는 것으로 해석되지는 않는다.

HNF-1α (Hepatocyte nuclear factor-1α) 유전자에서 알려진 돌연변이를 검출하기 위한 DNA 칩을 개발하는 과정에서 스폿팅(spotting) 방법에 의해 제작된 DNA 칩과 정상 표적과 혼성화했을 때와 돌연변이유발(mutagenesis)에

의해 합성된 돌연변이 염기서열을 가진 표적과 혼성화했을 때의 실험결과를 제시하고자 한다.

각 데이터는 20장의 칩에서 혼성화를 통해 실험결과를 얻었다. 한 장의 칩에는 각 돌연변이 위치당 정상형 프로브 2개(WP1 (wild type probe 1), WP2 (wild type probe 2)), 돌연변이형 프로브 2개(MP1(mutant type probe 1), MP2(mutant type probe 2))로 모두 4개씩의 프로브를 기판 위에 올려놓았다. 하기 표 1은 그 중에서 이후의 설명을 위해 몇몇 돌연변이 위치 (영문 대문자로 표기) 의 프로브 정보를 정리한 것이다.

DNA 칩 상에 프로브를 부착하는 과정은 다음과 같다. 하기 표 1에서와 같은 서열의 올리고뉴클레오타이드 프로브를 젤 매트릭스 용액에 첨가하고 교반하여 37℃에서 14시간 동안 방치함으로써 매트릭스-DNA 접합체(conjugates)를 제조한 다음 이를 스폿팅 용액으로 하였다(참조 : 대한민국 특허출원 제 2001-53687호). 상기 스폿팅 용액을 아민기를 갖도록 표면처리된 유리 표면 위에 스폿팅한 후 4시간 동안 37℃의 습식 챔버(wet chamber)에 방치하였다. 이어, 배경 노이즈의 제어(background noise control)에 필요한 공정, 즉 표적 핵산이 유리 표면에 부착하지 않도록 하기 위해 스폿팅 되지 않은 위치의 유리 표면 아민기가 음전하를 띄도록 반응을 실행하고 건조기에 보관하였다.

표1. 돌연변이 위치들의 프로브 샘플과 p 값

돌연변이 위치	프로브 타입	프로브 서열	p-value
3306	WP1	gaca C gcacctccgt	5.02454e-6
3306	MP1	tgtagaca A gcacct	5.56682e-3
3306	WP2	aca C gcacctccgtg	9.15422e-6
3306	MP2	tgtagaca A gcacct	3.11200e-3
4037	WP1	tgagacc T acgaggg	1.89666e-2
4037	MP1	ctgagacc G acgagg	2.82660e-4
4037	WP2	ctgagacc T acgaggg	1.04213e-1
4037	MP2	ctgagacc G acgagg	1.88030e-4
5683	WP1	ccac C ggctcagc	4.22692e-6
5683	MP1	cgctgagcc C gtgg	5.93080e-13
5683	WP2	ccac C ggctcagc	9.95383e-8
5683	MP2	gcgctgagcc C gtgg	5.25110e-12
6195	WP1	catcgaga C cttcatc	3.561e-6
6195	MP1	gatgaag A tctcgat	1.7029e-10
6195	WP2	cgtcatcgaga C cttc	5.0965e-11
6195	Mp2	gatgaag A tctcgat	9.3977e-11

실시예1> 돌연변이 위치 3306에 대한 혼성화 실험

돌연변이 위치 3306은 HNF-1 유전자 Exon4의 3306번 염기에 돌연변이가 초래되어 G가 T로 바뀐 것이다. 3306번 염기에 T가 존재하는 돌연변이형 표적에 상보적인 MP1과 MP2 프로브는 염기서열이 동일하다. 같은 염기서열이 한 칩에 두 번 탑재되었지만 평균차 검정 결과 얻어진 p 값은 5.567e-3과 3.11 2e-3으로 동일한 샘플에서 얻어지는 p 값이 크게 차이가 나지 않는다.

도4는 돌연변이 위치 3306의 두 돌연변이형 프로브(MP1, MP2)와 정상형/돌연변이형 표적과의 혼성화 강도 분포이다. 도4에서 한 점은 DNA 칩 한 장과 한 표적과의 혼성화 결과이다. X축은 표1에서 돌연변이 위치 3306이고 프로브 형이 MP1에 해당하는 올리고뉴클레오타이드 프로브와 3306에 G가 있는 정상형 표적과의 혼성화 강도를 파란색으로 표시하고, 3306에 T가 있는 돌연변이형 표적과의 혼성화 강도를 빨간 색으로 표시한 것이다. 마찬가지로 Y축은 표1에서 돌연변이 위치 3306이고 프로브 형이 MP2에 해당하는 프로브와 정상형 표적과의 혼성화 강도는 파란색으로, 돌연변이형 표적과의 혼성화 강도는 빨간 색으로 표시하였다. 따라서 파란색 점 중에서 가장 좌하에 위치한 점은 그 칩에서 정상형 표적과 MP1과의 혼성화 강도가 대략 240이고 MP2와의 혼성화 강도가 200가량 나왔다는 것을 의미한다. 그림에서 알 수 있듯이 MP1과 MP2의 염기서열로는 정상형 표적과 돌연변이형 표적과의 혼성화 강도 분포가 거의 겹치기 때문에 두 표적을 구분하기 어렵다. 이 때의 MP1과 MP2의 p 값은 각각 5.56682e-3과 3.11200e-3이

다.

실시예2> 돌연변이 위치 4037에 대한 혼성화 실험

돌연변이 위치 4037은 HNF-1 유전자 Intron5의 4037번 염기에 돌연변이가 초래되어 A가 G로 바뀐 것이다. 돌연변이 위치 4037에서는 돌연변이 위치 3306과 같이 돌연변이형 프로브와 돌연변이형 표적의 혼성화 강도가 정상형 표적과의 혼성화 강도와 겹치는 부분도 일부분(대략 1/4) 존재하지만 다수의 프로브 결과는 표적이 다름에 따라 혼성화 강도가 차이를 나타낸다. MP1과 정상/돌연변이형 표적과의 혼성화 강도를 기준으로 했을 때 8000이상이 되는 것을 돌연변이형 표적으로 정의한다면 2개의 정상형 표적을 돌연변이형으로 잘못 판별하고 5개의 돌연변이형 표적을 정상형으로 잘못 판별하지만 돌연변이 위치 3306에 대한 돌연변이형 프로브보다는 훨씬 더 정상형과 돌연변이형을 잘 판별한다. 이 때의 MP1과 MP2의 p 값은 각각 2.8226e-4과 1.8803e-4이다.

돌연변이형 표적 3306과 4037의 경우 뿐만 아니라 나머지 돌연변이 위치들에 대한 혼성화 강도의 분포를 조사한 결과, 정상형 표적과 돌연변이형 표적을 구분할 수 있는 프로브는 가장 큰 p 값이 1.00000e-3 보다 작음을 확인하였다. 따라서, 생물 / 의약 분야에서는 보통 유의수준으로 0.01를 잡지만 본 방법에서는 정확도를 높이기 위하여 0.001를 잡고서, 평균차 검정을 통해 얻어진 p 값이 1.00000e-3보다 작은 프로브를 유효한 프로브라고 정의한다.

실시예3> 돌연변이 위치 5683에 대한 혼성화 실험

돌연변이 위치 5683은 HNF-1 유전자 Exon9의 5683번 염기에 돌연변이가 초래되어 C가 G로 바뀐 것이다. 평균차 검정결과 MP1과 MP2에 의한 p 값은 각각 5.9308e-13과 5.2511e-12으로 유효한 프로브이지만 돌연변이형 프로브와 돌연변이형 표적과의 혼성화 강도가 정상표적과의 혼성화 강도보다 커야 한다는 당연한 사실을 위배한다. 이와 같이 완전 일치(perfect match: 돌연변이형 프로브-돌연변이형 표적)에 해당하는 혼성화 강도의 평균이 불일치(mismatch: 돌연변이형 프로브-정상형 표적)에 해당하는 혼성화 강도의 평균보다 작은 프로브를 잘못된 프로브(wrong probe)라고 정의한다.

실시예4> 돌연변이 위치 6195에 대한 혼성화 실험

돌연변이 위치 6195는 HNF-1 유전자 Exon10의 6195번 염기에 돌연변이가 초래되어 C가 T로 바뀐 것이다. 도면에서 볼 수 있듯이 완전 일치(돌연변이형 프로브-돌연변이형 표적)에 해당하는 혼성화 강도의 평균이 불일치(돌연변이형 프로브-정상형 표적)에 해당하는 혼성화 강도의 평균보다 크기 때문에 MP1과 MP2 프로브는 올바른 프로브이고 평균차 검정결과도 MP1과 MP2에 의한 p 값은 각각 1.7029e-10과 9.3977e-11으로 유효한 프로브이다. 우리가 컷오프(cuf-off) 기준으로 삼은 1.00000e-3보다 매우 작은 값을 가지고 있는데 이렇게 작은 p 값을 가지는 프로브 는 대체로 정상형 표적과 돌연변이형 표적의 분포가 완전히 분리되는 것을 볼 수 있다. 따라서 이러한 프로브를 심는다면 p 값이 가장 작은 프로브 하나만으로도 정상형 표적과 돌연변이형 표적을 판별할 수도 있고 복수의 프로브를 선택할 때에도 올바른 프로브들 중에서 p 값이 작은 순서로 선택하면 실험오차를 감안하더라도 유전자형 판별(genotyping)이 가능하다.

발명의 효과

이상 설명한 바와 같이, 본 발명에 따르면 정상 표적 핵산과 돌연변이 표적 핵산의 차이를 실험 오차 이내에서 감지할 수 있고, 점돌연변이 (point mutation) 뿐만 아니라 삽입(insertion) / 결실(deletion)의 경우에도 소수의 프로브를 가지고 유전자형을 확인할 수 할 수 있다.

(57) 청구의 범위

청구항 1.

(a) 정상 표적 핵산에 상보적인 프로브 및 돌연변이 표적 핵산에 상보적인 프로브가 고정된 바이오칩 상에서, 정상 표적 핵산 및 돌연변이 표적 핵산을 혼성화 (hybridization) 시켜 각 프로브별로 정상 표적 핵산-프로브 및 돌연변이 표적 핵산-프로브의 혼성화 강도 데이터를 수집하는 단계;

(b) 상기의 데이터를 이용하여 평균차 검정을 실시하여 유의한 차이가 나지 않는 프로브(p 값이 유의수준보다 큰 프로브)를 제외하는 단계;

(c) 유의한 차이를 가지는 프로브(p 값이 유의수준보다 작은 프로브)에 대하여 정상 표적 핵산-프로브 혼성화 강도 분포 및 돌연변이 표적 핵산-프로브 혼성 화 강도 분포로부터, 교차 유효성 검증법(cross-validation)에 의해 정상 표적 핵산을 돌연변이 표적으로 분류할 확률인 가짜 양성 에러율(false positive error rate) 또는 돌연변이 표적을 정상 표적 핵산으로 잘못 분류할 확률인 가짜 음성 에러율(false negative error rate)을 계산하는 단계; 및,

(d) 상기 가짜 양성 에러율 또는 가짜 음성 에러율의 기준을 통과한 프로브를 선택하는 단계를 포함하는 유전자형 확인용 프로브 세트 선택 방법.

청구항 2.
제1항에 있어서, 상기 (b)단계는

(a) 두 표본(정상 표적 핵산을 가지는 샘플-프로브와의 혼성화 강도 분포, 돌연변이 표적 핵산을 가지는 샘플-프로브와의 혼성화 강도 분포)에 대해서 모두 정규분포를 따르면 두 표본 t-검정(two sample t-test)을 실시하여 p 값을 구하고, 어느 하나라도 정규분포를 따르지 않는 경우는 비모수적 방법(nonparametric method)에 의해 p 값을 구하는 단계;

(b) 상기 p 값으로 두 표본의 평균이 유의한 차이가 있는가(p값이 유의수준보다 작은가)를 확인하는 단계;

(c) 상기 (b)단계에서 유의한 차이가 나는 프로브에 대해 정상형 프로브의 경우는 정상 표적 핵산과의 혼성화 강도 평균이 돌연변이 표적 핵산과의 혼성화 강도 평균보다 크고 돌연변이형 프로브는 반대로 돌연변이 표적 핵산과의 혼성화 강도 평균이 정상 표적 핵산과의 혼성화 강도 평균보다 큰 프로브를 선발하는 단계를 포함하는 것이 특징인 유전자형 확인용 프로브 세트 선택 방법.

청구항 3.
제2항에 있어서, 상기 (a)단계에서 두 표본 t-검정을 실시하기 전에 각 샘플에 대한 혼성화 강도의 분산이 같은지를 검정(등분산성 검정)하여 같은 경우는 t-검정에서 나온 결과 중에 등분산에 해당하는 p 값을 선택하고, 다른 경우는 t-검정에서 나온 결과 중에 등분산이 아닌 경우에 해당하는 p 값을 선택하는 단계를 더 포함하는 것이 특징인 유전자형 확인용 프로브 세트 선택 방법.

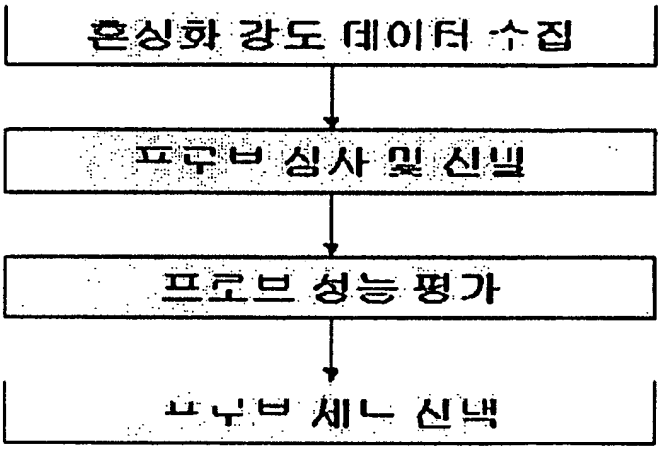
청구항 4.
제2항에 있어서, 상기 (b)단계에서의 유의수준은 0.01인 것이 특징인 유전자형 확인용 프로브 세트 선택 방법.

청구항 5.
제 2 항에 있어서, 상기 (b)단계에서의 유의수준은 0.0 0 1인 것이 특징인 유전자형 확인용 프로브 세트 선택 방법.

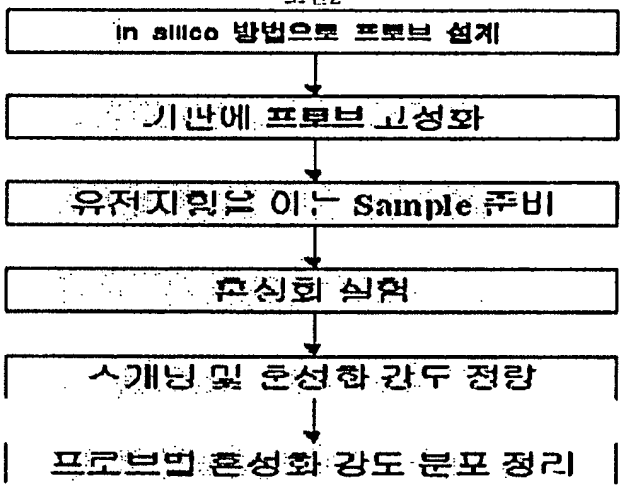
청구항 6.
제 2 항에 있어서, 표적 핵산이 HNF-1α (Hepatocyte nuclear factor-1α) 유전자인 경우에는, 상기 (b)단계에서의 유의수준은 0.001인 것이 특징인 유전자형 확인용 프로브 세트 선택 방법.

도면

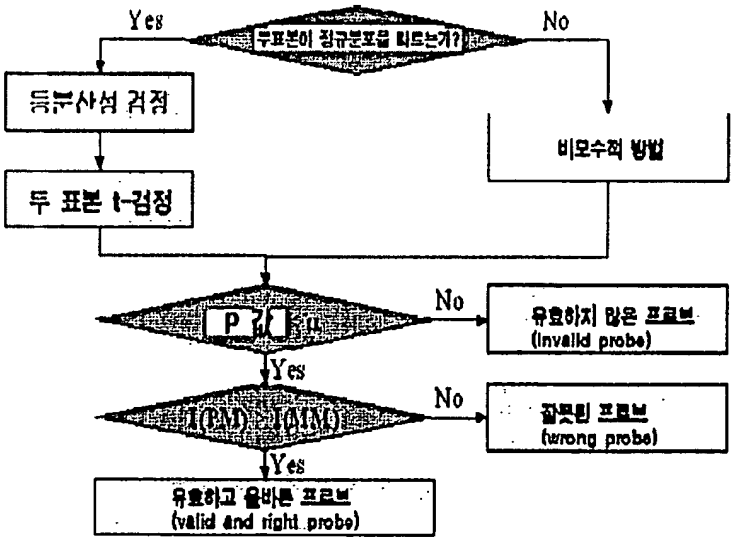
도면1


혼성화 강도 데이터 수집
프로브 설사 및 신빌
프로브 성능 평가
프로브 세트 선택

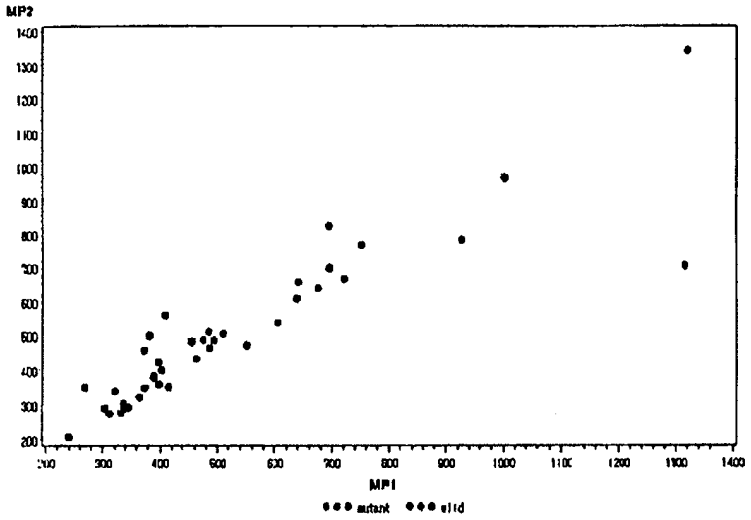
도면2


in silico 방법으로 프로브 설계
기판에 프로브 고정화
유전자형을 이는 Sample 준비
혼성화 실험
스캐닝 및 혼성화 강도 정량
프로브별 혼성화 강도 분포 정리

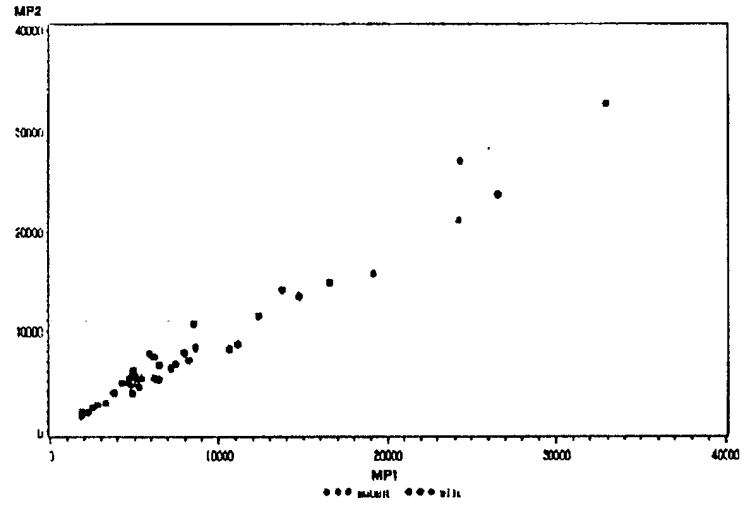
도면3


Yes
No
두표본이 정규분포를 따르는가?
등분산성 검정
두 표본 t-검정
비모수적 방법
No
유효하지 않은 프로브
(invalid probe)
Yes
No
잘못된 프로브
(wrong probe)
Yes
유효하고 올바른 프로브
(valid and right probe)

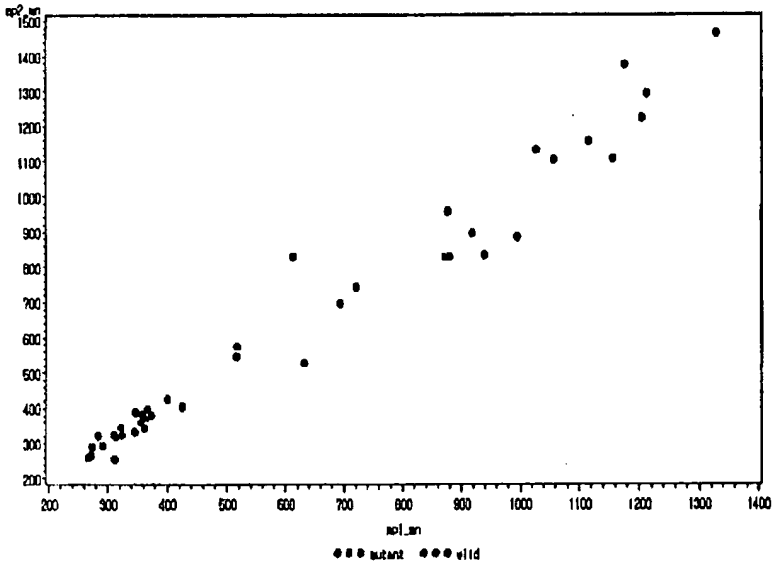
도면4


MP2
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
MP1
mutant
wild

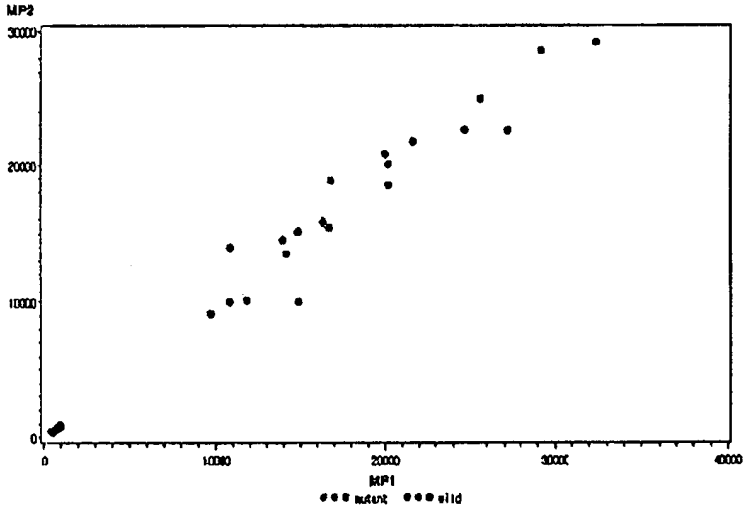
도면5


MP2
40000
30000
20000
10000
0
0
10000
20000
30000
40000
MP1

도면6


mp2_an
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
mp1_an
mutant
wild

도면7


MP2
30000
20000
10000
0
0
10000
20000
30000
40000
MP1
mutant
wild